Getting Off the "Gold Standard": Randomized Controlled Trials and Education Research

Gail M. Sullivan, MD, MPH

Randomization may be achieved at the expense of relevance.

L. J. Cronbach, *Designing Evaluations of Educational and Social Problems*[1]

The Randomized Controlled Trial "Gold Standard"

Medical education researchers are inevitably more familiar with the multisite randomized controlled trial (RCT) "gold standard" from the clinical research paradigm. In an RCT, trainees are randomly assigned to receive 1 of 2 or more educational interventions. Randomized controlled trials are quantitative, comparative, controlled experiments in which treatment effect sizes may be determined with less bias than observational trials.[2] Randomization is considered the most powerful experimental design in clinical trials: with other variables equal between groups, on average, any differences in outcome can be attributed to the intervention.[2]

With trainees moving through educational processes in real time, the difficulties of randomization become immediately clear. Residents and fellows usually experience rotations at different times: those experiencing the intervention later may learn from intervening experiences and no longer be comparable to those experiencing the intervention at an earlier time. Also, trainees have the option to refuse randomization, yet they cannot miss critical educational experiences. Use of a "placebo" is often contraindicated. Multisite interventions, while providing more subjects (with greater power to detect differences) and more generalizability, present challenges due to training differences. Also, the cost of multisite studies can be high and industry support for education trials rarely exists.

While large numbers of medical undergraduates often receive nearly identical "treatments," in which 1 or more variables may be altered, this is not true for graduate medical education. Residency and fellowship training are usually highly individualized, which makes the RCT model increasingly unsuitable as training advances.

Gail M. Sullivan, MD, MPH, is Editor-in-Chief, *Journal of Graduate Medical Education.*

Corresponding author: Gail M. Sullivan, MD, MPH, University of Connecticut, 253 Farmington Avenue, Farmington, CT 06030-5215, gsullivan@nso1.uchc.edu

DOI: http://dx.doi.org/10.4300/JGME-D-11-00147.1

In addition to feasibility considerations, experts question the applicability of research models, derived from clinical research, for education studies.[3,4] The highly complex system of education may be a poor fit for the RCT model, which requires clear inclusion/exclusion criteria and interventions administered identically via multiple physicians (ie, teachers).[3,5,6] Regehr asks whether simulating placebo-controlled efficacy clinical trials, in which 1 or a few variables may be tightly controlled, is a worthwhile goal for medical education research.[3] In education studies, variables can rarely be controlled tightly and blinding of subjects and study personnel may be unethical or impossible.[3,7] Finally, defining the therapeutic intervention in education research is much more difficult than in clinical trials. As Norman suggests, one cannot "apply curriculum daily" in the same way that one can prescribe a medication.[5]

RCTs in Education Research

The primary advantage of randomization is that it reduces allocation bias, which derives from baseline variables that may influence outcome(s). Randomization ensures that baseline characteristics, not known to be related to the outcome of interest, are equally distributed among the groups. Although differences among participants are 1 source of error, *randomization will not control for other sources of error, which are likely to occur in education studies* (TABLE 1). Variations in those implementing the intervention, settings, and other execution factors may have more impact than baseline variations in the subjects. Other common "confounders" that randomization may not control for are effects of pretests on learning (encouraging differential study or learning);[8] Hawthorne effects (changes in participant motivation); effects of other, nonintervention training experiences occurring during the study intervention;[5] and high participant dropout (eg, less than 75% response rates).[9] Contextual factors may affect outcomes in ways that randomization cannot fix.[10] Especially if the intervention is fairly dilute (eg, a workshop, short course, or online cases), it may not be apparent whether the intervention is causing the outcome effects versus contextual factors.

In education studies it is often difficult to "blind" learners to their assigned group. Without blinding, residents can react to the knowledge that they are being studied or assigned to a particular group. Within training programs trainees interact to a great extent, resulting in contamination effects (ie, trainees sharing learning with

TABLE 1 **Sources of Error in Experimental Education Studies**[8,14]

Source of Error	Solutions
Different sites of intervention	Explicitly describe sources of variability and potential effects on outcomes
Students at different times in training	Concurrent comparison group
Different implementers (teachers)	Training manual; enumerate and describe variations explicitly
Difference in intensity of intervention	Describe variability (eg, trainees miss intervention owing to patient care responsibilities, vacation, etc.) and potential effects on outcomes
Low response rate or high dropout rate	Describe characteristics of responders and nonresponders in as much detail as possible
No concurrent comparison group	Nonconcurrent comparison group; describe variables that may be different between the groups and potential effects (in both directions) upon outcomes
Use of "historical" controls	Descriptions of the characteristics of the intervention and historical group subjects, as well as detailed description of previous teaching activities
	This design may preclude firm conclusions about the intervention
No control group (*eg, single group pretest/posttest design*)	Results can suggest hypotheses for further testing with a stronger design
	Multiple sources of error preclude firm conclusions

each other) that further compromise randomization. Active interventions that are deemed critical to learning cannot be withheld, although crossover designs may be used in this situation. However, crossover designs may also involve contamination of learning between groups.

When should an RCT be used in education experiments? According to Norman,[6] randomization is most useful in examining relatively standardized interventions, such as web-based learning and, possibly, clinical simulation. He recommends that randomization be considered when (1) prior observational studies support the hypothesis; (2) the mechanism of learning is understood; (3) the outcome of the intervention is easily measured and accepted as related to the intervention; (4) the subgroups likely to benefit from the intervention are also easily identified; (5) the effect size of the intervention is small; and (6) the results from the trial may have a large impact, to justify the costs of an RCT[6] (TABLE 2). These criteria are not often satisfied in medical education studies.

Even in clinical research, RCTs are most helpful for therapeutic trials, rather than for risk factor identification or prognosis.[6] Likewise in education research, there are research questions for which randomization will be inappropriate: residents cannot be randomly assigned to whether they are from rural versus urban areas, married, or female. Studies of predictive factors and career choices will need cohort, case-control, and case-series designs. In summary, randomization in medical education research is not a cure and not the best method for many research hypotheses.

Alternatives to Randomization

While qualitative research is an obvious starting point for many educational questions, this valuable method will be discussed in a future editorial, while we focus here on quantitative methods (TABLE 3). Nonrandomized methods are common in education research and considered by experts as not inferior to RCTs. In systematic reviews, Best Evidence in Medical Education groups grade the strength of articles on several factors, but not whether the study was randomized.[11]

Perhaps a more relevant clinical research model for educators is the "pragmatic trial."[12] In a pragmatic trial, 2 or more medical interventions are compared in real-world practice. Patients are heterogeneous from a wide variety of practice settings, nonblinded, and may choose to switch treatments. Of note, nonadherence, a major threat to useful findings from pragmatic clinical trials, may not be as problematic in educational studies. However, a much greater number of subjects are usually needed to determine true differences (or equivalence) among interventions, which will present a challenge for education researchers, as the structure and funds for large multi-institution studies are at present scarce. The pragmatic trial paradigm has been suggested for patient safety research, in which context is also a critical variable.[13] As considerable overlap exists, examining successful research designs in patient safety research may yield insights for education research.

Borrowing from epidemiology research methods, observational designs can be cross-sectional or longitudinal. Longitudinal studies may use ongoing surveillance or repeated cross-sectional methods to measure change over time.[14] To strengthen these research designs, one must include a comparison group. The comparison group may be a concurrent cohort or matched "case-control" format. These studies may be prospective (assembled, described, and

TABLE 2 **Randomized Controlled Trials in Clinical Research Versus Education Research[6]**

Characteristic	Clinical Example	Education Studies
Mechanism of action well understood	Aspirin to prevent MI	Learning theories may guide research and be considered in the planning phase; often mechanism of learning not well understood
Endpoint easily measured	MI clinical criteria	Diverse outcomes: knowledge; behaviors with simulation; behaviors with various people/patients; attitudes regarding importance
Targets, likely to benefit from intervention, easily understood	Risk factors for MI (or aspirin toxicity) used to create inclusion criteria	Individuals may respond differently to different educational approaches: identification of these individuals generally not feasible
Effect size of intervention expected to be small	0.07 = effect size for aspirin preventing MI	0.5 effect size, on average, in 1 systematic review of successful interventions[4] (ie, this is large)
Effect of intervention independent of who is administering	"81-mg aspirin" daily is not dependent on physician attributes	Teacher characteristics and teacher-student interactions may greatly affect success of intervention
Major impact if trial is successful	Large population effect of reducing MIs; large pharmaceutical company profits from new, effective drugs	Little profit in medical education; successful interventions often remain local

Abbreviation: MI, myocardial infarction.

followed forward) or retrospective, if sufficient past data on key variables are available.

One of the most common designs noted in original research submissions to the *Journal of Graduate Medical Education* is the single group design. Without a comparison group, this design may suggest hypotheses for future study, but will not generate firm conclusions. Potential alternatives include the use of historical controls—with inherent

TABLE 3 **Choosing a Quantitative Education Research Design**

Research Design	Benefits and Disadvantages
Randomization	Random distribution of subject characteristics in test groups
	Random likelihood of changes occurring to subjects during period of study
Comparison group	Equal exposure to changes in ascertainment or scoring of outcomes
Concurrent comparison group	Equal exposure to intervening events, not part of intervention, that may impact outcomes
Pretest/posttest (where pretest score is used as a covariate in multivariate analysis[2])	Useful for nonrandomized studies or randomly assigning less than 40 subjects[19]
	Useful if high dropout rate likely[8]
	May introduce error as students, cued to the study hypothesis from the pretest, study for the posttest or are more motivated to learn—*discuss in limitations*
Posttest only	Avoids pretest highlighting study hypothesis for subjects
	Use if randomly assigning at least 40 subjects[19]
Blinding of participants to research hypothesis	Reduces differential learning due to subjects
	1) aware they are being studied and
	2) motivated to study or learn differently in active versus control groups
Blinding of teachers to research hypothesis	Equal implementation of interventions: difficult to achieve
	Alternative: teachers equally well trained to deliver educational intervention—*describe in methods*
Paired analysis of pretest/posttest (ie, subtract pretest from posttest)	Controls for differing baselines in subjects
	Potentially doubles error (use if reliability coefficient, for each test, is sufficiently high)—*describe test reliability in methods*

BOX Suggestions to Strengthen Quantitative Research Methods

- Adequate sample size
  - More than 1 iteration of the intervention
  - Multiple sites
  - Low dropout rates/high participation rates
- Comparison group—*describe thoroughly*
  - Crossover
  - Historical
  - Different site with usual teaching, without new intervention
- Comparison group receives an active intervention
- Residents not cued to "new" approach
  - No pretest unless already part of rotation/experience
- Ensure equal application of intervention
  - Training manual or other method, for teachers
- Rich discussion of potential bias in study
  - Use limitations to fully explore alternatives to original hypothesis
- Next steps: more rigorous methods to confirm findings[1]

concerns about significantly different cohorts—and crossover alternate rotation designs. Residents may be assigned to the intervention on alternate rotations, with full discussion of potential bias in the assignments.[15] A repeated measures crossover design is valuable in many nonrotation-based interventions as well.

Validity Concerns With RCT and Non-RCT Research

When the priority is to find and publish positive results, less consideration may be given to the causes of the differences observed.[3] The key question is whether the differences are due to the intervention or to potential bias. Experts assert that often it is confounders that cause the positive results in education studies, rather than the intervention itself.[3,16] If these confounding factors are discussed thoroughly, important insights may result and actually provide more enlightenment than the ''positive findings.'' When nonrandomized, noncontrolled designs are used, Colliver and McGaghie emphasize that the potential ''threats to validity'' thus introduced must be discussed thoroughly in ''a central place in the study'' rather than as a perfunctory list in the limitations section.[16] Without this meticulous analysis of potential confounding variables, important research questions are missed and overinterpretation of results is common.

Researchers are often faced with the problem of small sample sizes. Several iterations of the intervention and data collection may be necessary to obtain sufficient numbers of subjects, for firm conclusions. In research submissions to the *Journal of Graduate Medical Education*, this potential solution is often overlooked. Some researchers, faced with a small number of subjects to study, may forgo a control group: all subjects receive the intervention. These are often termed ''show and tell'' studies, reports of a single iteration that happens to find a positive result.

While randomization is not necessary, a comparison group *is* essential in education research. Sometimes teachers initiate curriculum improvement or new mandated requirements first and later decide to publish as a research study.[17] If the intervention involved all available subjects, a delayed search for a comparison group may be difficult, yet not impossible. Any comparison group is better than none, but dissimilar comparison groups may introduce a large degree of bias. While these descriptive studies may generate useful hypotheses and stimulate key discussions, more rigorous methods will be needed to build evidence in favor of the intervention and should be the ''next step'' for alert researchers.

Summary

While useful in some situations, randomization is not the ''gold standard'' for medical education research. More important is that decisions regarding methodology precede the intervention, that adequate numbers of subjects and iterations are used, that a comparison group is included, and that limitations are addressed in a thoughtful, thorough manner.

In addition, the literature demonstrates quite definitively that medical learners will learn whatever we teach and also may supplement any teaching deficits to meet certification requirements.[18] Thus, to increase our understanding of effective educational interventions, a new educational intervention should be compared to another effective intervention. Unlike clinical research, a placebo arm is rarely helpful. Comparing the new educational intervention to ''usual'' practices is productive as long as students are not ''cued'' to the novelty of the research arm—which may enhance (or negatively bias) their learning—and the usual practices are well described.

Whether randomized or nonrandomized, medical education studies must carefully analyze sources of bias—unforeseen confounding variables—to explain the observed results and ensure that subsequent researchers will find these results reproducible. Rather than an obligatory listing of these potential sources of error in the discussion, researchers will enhance existing knowledge through a careful and detailed analysis of sources of bias that may have affected the results.[3] Equally important, research designs must include a control group, concurrent if possible, to ensure equal likelihood of exposure to nonintervention events that could bias the results.

If educators begin to work together, more collaborative, multi-institutional projects, perhaps akin to ''pragmatic trials,''[12] may be produced in the future. This is likely to add substantially to our understanding of effective resident education.

References

1 Cronbach LJ. *Designing Evaluations of Educational and Social Problems*. San Francisco, CA: Jossey-Bass; 1982.

2 Stolberg HO, Norman G, Trop I. Fundamentals of clinical research for radiologists: randomized controlled trials. *Am J Roentgenol*. 2004;183:1539–1544.

3 Regehr G. It's NOT rocket science: rethinking our metaphors for research in health professions education. *Med Educ*. 2010;44:31–39.

4 Eva KW. Broadening the debate about quality in medical education research. *Med Educ*. 2009;43:294–296.
5 Norman G. RCT = results confounded and trivial: the perils of grand educational experiments. *Med Educ*. 2003;37:582–584.
6 Norman G. Is experimental research passé. *Adv Health Sci Educ Theory Pract*. 2010;15(3):297–301.
7 Prideaux D. Researching the outcomes of educational interventions: a matter of design. *BMJ*. 324:126–127.
8 Cook DA, Beckman TJ. Reflections on experimental research in medical education. *Adv Health Sci Educ Theory Pract*. 2010;15(3):455–464.
9 Reed DA, Beckman TJ, Wright SM, Levine RB, Kern DE, Cook DA. Predictive validity evidence for medical education research study quality instrument scores: quality of submissions to JGIM's medical education special issue. *J Gen Intern Med*. 2008;23:903–907.
10 Berliner DC. Educational research: the hardest science of all. *Educ Researcher*. 2002;31:18–20.
11 Harden RM, Grant J, Buckley G, Hart IR. BEME guide No. 1: best evidence medical education. *Med Teach*. 1999;21:553–562.
12 Ware JH, Hamel MB. Pragmatic trials—guides to better patient care? *New Engl J Med*. 2011;364:1685–1687.
13 Clancy CM, Berwick DM. The science of safety improvement: learning while doing. *Ann Intern Med*. 2011;154:699–701.
14 Carney PA, Nierenberg DW, Pipas CF, et al. Educational epidemiology: applying population-based design and analytic approaches to study medical education. *JAMA*. 2004;292:1044–1050.
15 Shea JA, Arnold L, Mann KV. RIME perspective on the quality and relevance of current and future medical education research. *Acad Med*. 2004;79:931–938.
16 Colliver JA, McGaghie AC. The reputation of medical education research: quasi-experimentation and unresolved threats to validity. *Teach Learn Med*. 2008;20:101–103.
17 Gruppen LD. Improving medical education research. *Teach Learn Med*. 2007;19:331–335.
18 ten Cate O. What happens to the student: the neglected variable in educational outcome research. *Adv Health Sci Educ Theory Pract*. 2001;6(1):81–88.
19 Fraenkel JR, Wallen NE. *How to Design and Evaluate Research in Education*. New York, NY: McGraw-Hill; 2003.